Collective Interpretation and Potential Joint Information Maximization

Ryotaro Kamimura

Collective Interpretation and Potential Joint Information Maximization

Ryotaro Kamimura

IT Education Center and Graduate School of Science and Technology
Tokai Univerisity 4-1-1 Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan
ryo@keyaki.cc.u-tokai.ac.jp

Abstract. The present paper aims to propose a new type of information-theoretic method called "potential joint information maximization". The joint information maximization has an effect to reduce the number of jointly fired neurons and then to stabilize the production of final representations. Then, the final connection weights are collectively interpreted by averaging weights produced by different data sets. The method was applied to the data set of rebel participation among youths. The result show that final weights could be collectively interpreted and only one feature could be extracted. In addition, generalization performance could be improved.

Key words: Collective interpretation, generalization, mutual information maximization, potentiality, pseudo-potentiality

1 Introduction

Information-theoretic methods have had much influences on neural computing in many aspects of neural learning [1], [2], [3], [4], [5], [6], [7]. Though the information-theoretic methods have aimed to describe relations or dependencies between neurons or between layers, due attention has not been paid to those relations. They have even tried to reduce the strength of relations between neurons [8], [9]. For example, they have tried to make individual neurons as independent as possible. In addition, they have tried to make the distribution of neurons' firing as uniform as possible. This is simply because difficulty has existed in taking into account neurons' relations or dependencies.

The present paper aims to describe one of the main relations between neurons, namely, relations between input and hidden neurons, because they play critical roles in improving the performance of neural networks, for example, generalization performance. However, it has been few efforts to describe relations between input and hidden neurons from the information-theoretic points of view. To examine relations between input and hidden neurons, we introduce the joint probability between input and hidden neurons. Then, the joint information contained between input and hidden neurons is also introduced. When this joint information increases, only a small number of joint input and hidden neurons fire strongly, while all the others cease to do so.

However, one of the major problems to realize the joint information lies in difficulty in computation. As has been well known, the majority of the information-theoretic methods have this problem of difficulty in computation [7]. To overcome the problem, we have introduced the potential learning [10], [11], [12], [13]. In the method, information maximization can be translated into potentiality maximization where a specific neuron is forced to have the largest potentiality to deal with many different situations. Applying the potentiality to joint neurons, potentiality maximization corresponds to a situation where a small number of joint neurons are forced to have larger potentiality.

In addition, the present method aims to propose a new method to interpret final representations. As has been well known, the black-box property of neural networks have prevented them from being applied to practical problems, because in practical applications, the interpretation of final results can be more important than the generalization performance. Usually, neural networks produce completely different types of connection weights, depending on different data sets and initial conditions. The joint information maximization can be used to explain the final representations clearly. When the joint information increases, the number of activated neurons diminishes, which constraints severally the production of many different types of weights. Thus, a few typical connection weights are produced by the joint information maximization. Then, we can interpret those connection weights by averaging them. This type of interpretation is called "collective interpretation" in the present paper. As generalization performance is evaluated in terms of the average values, the interpretation performance can be evaluated collectivity by taking into account all the connection weights produced by diffident data sets and initial conditions.

2 Theory and Computational Methods

2.1 Concept of Joint Information Maximization

Figure 1 shows a concept of joint information maximization. For a data set, when the joint information is maximized, only one joint hidden and input neuron fire strongly with a strong connection weight in Figure 1(b). For another data set, another joint hidden and input neuron strongly fire in Figure 1(c). For interpretation, connection weights produced by all data sets are taken into account by averaging connection weights with due consideration for hidden-output connection weights in Figure 1(e).

2.2 Potential Joint Information Maximization

Potential joint information is based on the potentiality so far defined for hidden neurons [10], [11], [12], [13]. As shown in Figure 1(b), let w_{jk}^t denote connection weights from the kth input neuron to the jth hidden neuron for the tth data set, then the potentiality v_{jk}^t is defined by

$$v_{jk}^t = \left(w_{jk}^t - w^t\right)^2, \tag{1}$$

Fig. 1. Concept of joint information maximization with collective interpretation.

where w^t denotes the average weight defined by

$$w^t = \frac{1}{ML} \sum_{j=1}^{M} \sum_{k=1}^{L} w_{jk}^t, \tag{2}$$

where M and L denotes the number of hidden and input neurons. Then, the potentiality is normalized as

$$p(j,k|t) = \frac{v_{jk}^t}{\sum_{m=1}^{M} \sum_{l=1}^{L} v_{ml}^t}. \tag{3}$$

Then, we have the potential joint information

$$PJI = -\sum_{j=1}^{M} \sum_{k=1}^{L} p(j,k) \log p(j,k) + \sum_{t=1}^{T} p(t) \sum_{j=1}^{M} \sum_{k=1}^{L} p(j,k|t) \log p(j,k|t), \tag{4}$$

where T is the number of data sets, $p(t)$ is the probability with which the tth data set is given and

$$p(j,k) = \sum_{t=1}^{T} p(t)p(j,k|t). \tag{5}$$

2.3 Computing Pseudo-Potential Joint Information Maximization

It is possible to differentiate the joint information to have update rules, but much simpler methods have been developed in the name of potential learning. In the method, potentiality maximization is replaced by pseudo-potentiality maximization, which is easily maximized just by changing the parameter. Now, the pseudo-potentiality is defined by

$$\phi_{jk}^{t,r} = \left(\frac{v_{jk}^t}{v_{max}^t} \right)^r, \tag{6}$$

where $r \geq 0$ deontes the potential parameter v_{max} is the maximum potentiality. By normalizing this potentiality, we have the pseudo-firing probability

$$p(j,k|t;r) = \frac{\phi_{jk}^{t,r}}{\sum_{m=1}^{M}\sum_{l=1}^{L}\phi_{ml}^{t,r}}. \tag{7}$$

Then, we have pseudo-information

$$\begin{aligned} PPJI = &- \sum_{j=1}^{M}\sum_{k=1}^{L} p(j,k;r)\log p(j,k;r) \\ &+ \sum_{t=1}^{T} p(t) \sum_{j=1}^{M}\sum_{k=1}^{L} p(j,k|t;r)\log p(j,k|t;r). \end{aligned} \tag{8}$$

The pseudo-information can be increased just by increasing the parameter r, and the joint information can be increased by assimilating pseudo-potentiality $\phi_{jk}^{t,r}$ repeatedly, while the potential parameter increased gradually. The new weights ${}^{new}w_{jk}^t$ are obtained by weighting the old weights ${}^{old}w_{jk}^t$ by the pseudo-potentiality

$${}^{new}w_{jk}^t = {}^{old}w_{jk}^t \, \phi_{jk}^{t,r}. \tag{9}$$

Then, new learning starts with those connection weights as initial ones. This process repeats itself for a fixed number of learning steps.

3 Results and Discussion

3.1 Experimental Outline

The data set was made to infer the probability of rebel participation among youths in the Niger Delta [14]. The number of input patterns was 1,340, and 19

Fig. 2. Potential joint information with 10 hidden neurons for the rebel data set.

input variables were used. The number of patterns for modeling neural networks was 1000 and the remaining 340 was exclusively for testing. With less than 1000 patterns, improved generalization performance was not obtained by the present and conventional methods. Of 1000 modeling data, 700 training data were randomly and repeatedly taken and ten training sets were prepared. The remaining 300 were used for the early stopping and checking the data sets. The potential parameter r was gradually increased from zero in the first learning step to one in the tenth learning step (final step).

3.2 Mutual Information

Figure 2 shows the joint information as a function of the number of steps. The joint information was simplified by supposing the uniform distribution

$$PJI = \log MN + \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{M}\sum_{k=1}^{L} p(j,k|t)\log p(j,k|t). \tag{10}$$

The information increased gradually and close to 0.6. Though the joint information could be further increased, generalization errors increased in direct proportion to this information increase beyond this point. The results show that the present method can increase the joint information sufficiently.

3.3 Connection Weights

Figure 3 shows connection weights for the rebel data set when the number of steps increased from one to ten. When the number of steps was one, almost random weights could be seen in Figure 3(a). When the number of steps was increased from two in Figure 3(b) to six in Figure 3(f), gradually the number of strong connection weights decreased. Then, when the number of steps was increased

Fig. 3. Connection weights from input to hidden neurons with 10 hidden neurons for the rebel data set. Green and red weights represent positive and negative ones.

from seven in Figure 3(g) to ten in Figure 3(j), only one connection weight from the eighth input neuron to sixth hidden neuron became the strongest, while all the other weights became close to zero.

Figure 4 shows adjusted connection weights for the maximum potential hidden neurons j^* by ten different data sets randomly taken from the modeling

data set. Adusted weights for interpretation $c^t_{j^*k}$ was computed by

$$c^t_{j^*k} = \text{sign}(W^t_{1j^*})w^t_{j^*k}, \tag{11}$$

where $\text{sign}(W_{1j*})$ denote the sign of the weight from the maximum potential hidden neuron to the first output neuron, representing that the youths do not want to participate in the rebel force. As shown in the figure, five out of ten results showed that the input neuron No.8 had stronger weights than any other ones. Thus, the input neuron No.8 was collectively considered to be important by the present method.

Figure 5 shows the average connection weights. The average weights were computed by

$$\bar{c}_{j*k} = \frac{1}{T}\sum_{t=1}^{T} c^t_{j^*k} \tag{12}$$

As can be seen in the figure, the input neuron No.8 had the largest connection weight. The variable No.8 represents the government's presence in the community in terms of the number of government establishments. Thus, when the government's presence becomes more visible, the youths do not want to participate in the rebel force.

Figure 6 shows the regression coefficients by the logistic regression analysis. In the original data set, a tricky variable was introduced, namely, the variable No.16 (oil size) and No.17 (squared oil size), which were naturally correlated, because principally two variables were the same. Thus, they produced the multi-collinearity where two variable responded completely differently to input patterns. On other hand, the present method responded to the two variables almost evenly. The results show that the present method is good at dealing with this kind of data set with strong correlation between variables. Finally, the interesting thing to note is that except the variables No.8, No.16 and No.17, quite similar weights and coefficients were produced by both methods.

3.4 Generalization Performance

The present method produced the best performance of generalization, comparing with that by the other two conventional methods. Table 1 shows generalization performance by three methods. As can be seen in the table, the best generalization error of 0.1662 on average was obtained by the present method. In addition, the best minimum and maximum error of 0.1382 and 0.2 were obtained by the present method. The second best one was obtained by the BP with the early stopping. Finally, the worst one was obtained by the logistic regression analysis.

4 Conclusion

The present paper proposed a new information-theoretic method called "joint information maximization". The joint information represents relations between

Fig. 4. Adjusted connection weights for ten different data sets from input to hidden neurons with 10 hidden neurons for the rebel data set. Green and red weights denote positive and negative ones.

input and hidden neurons. When the joint information increases, the number of strongly connected hidden and input neurons decreases gradually. The method

Fig. 5. Collective and average weights for the rebel data set.

Fig. 6. Regression coefficients for the rebel data set.

Table 1. Summary of experimental results on generalization performance for the rebel data set. The BP(ES) represents the BP with early stopping. The bold face numbers show the best values.

Method	Step	Hidden	Average	Std dev	Min	Max	Inf
Joint	6	10	**0.1662**	0.0181	**0.1382**	**0.2000**	0.4647
BP(ES)	1	10	0.1788	0.0338	**0.1382**	0.2529	0.1262
Logistic			0.2106	0.0129	0.1853	0.2294	

was applied to the rebel participation data set. The results show that the joint information could be increased by the present method. Final results could be interpreted collectively by averaging the connection weights. Finally, generalization performance was improved by the present method. The present method was much simpler than any other conventional information-theoretic methods because of the potential learning. Thus, it can be applied to large-scale and practical problems.

References

1. R. Linsker, "Self-organization in a perceptual network," *Computer*, vol. 21, no. 3, pp. 105–117, 1988.
2. H. B. Barlow, "Unsupervised learning," *Neural computation*, vol. 1, no. 3, pp. 295–311, 1989.
3. G. Deco, W. Finnoff, and H. Zimmermann, "Unsupervised mutual information criterion for elimination of overtraining in supervised multilayer networks," *Neural Computation*, vol. 7, no. 1, pp. 86–107, 1995.
4. A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.
5. R. Linsker, "Improved local learning rule for information maximization and related applications," *Neural networks*, vol. 18, no. 3, pp. 261–265, 2005.
6. J. C. Principe, D. Xu, and J. Fisher, "Information theoretic learning," *Unsupervised adaptive filtering*, vol. 1, pp. 265–319, 2000.
7. J. C. Principe, *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.
8. P. Comon, "Independent component analysis: a new concept," *Signal Processing*, vol. 36, pp. 287–314, 1994.
9. A. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.
10. R. Kamimura, "Self-organizing selective potentiality learning to detect important input neurons," in *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, pp. 1619–1626, IEEE, 2015.
11. R. Kamimura and R. Kitajima, "Selective potentiality maximization for input neuron selection in self-organizing maps," in *Neural Networks (IJCNN), 2015 International Joint Conference on*, pp. 1–8, IEEE, 2015.
12. R. Kamimura, "Supervised potentiality actualization learning for improving generalization performance," in *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, p. 616, The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2015.
13. R. Kitajima and R. Kamimura, "Simplifying potential learning by supposing maximum and minimum information for improved generalization and interpretation," in *Modelling, Identification and Control, 2015 International Conference on*, IASTED, 2015.
14. A. Oyefusi, "Oil and the probability of rebel participation among youths in the niger delta of nigeria," *Journal of Peace Research*, vol. 45, no. 4, pp. 539–555, 2008.